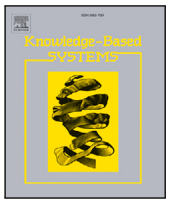
Knowledge-Based Systems 350 (2026) 116458

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Emotion-aware multimodal lightweight framework for adaptive voice interaction on edge device

Van-Duc Khuat [a], Jeongin Kim [b], Sang-Ho Kim [c], Yue Cao [d], Martin Maier [e], Wansu Lim [a,*]

[a] Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, Republic of Korea
[b] Division of Artificial Intelligence & Software, Ewha Womans University, Seoul, Republic of Korea
[c] Kumoh National Institute of Technology, Gumi, Republic of Korea
[d] School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China
[e] Optical Zeitgeist Laboratory, INRS, Montreal, QC H5A 1K6, Canada

ARTICLE INFO

Keywords:
Adaptive VUI
Multimodal emotion recognition
Personalized interaction
Model compression
Edge deployment

ABSTRACT

Voice user interfaces (VUIs) have been a major focus of human–computer interaction research over the past two decades. Advances in voice and emotion recognition have significantly enhanced their capabilities. However, VUIs still struggle to produce responses that fit the user's emotions in various contexts. Meanwhile, the growing demand for on-device intelligence makes efficient operation on resource-constrained edge devices increasingly important. To address these issues, this study presents an emotion-aware VUI framework designed for edge devices. The framework includes a lightweight multimodal emotion recognition model that integrates electroencephalogram signals and facial expressions. Furthermore, to enable on-device efficiency, this paper applies filter pruning to remove redundant filters. It uses cross-layer ranking together with k-reciprocal nearest-filter selection. This improves cross-layer consistency and preserves salient representations while sharply reducing computation. Post-training quantization is then applied to further shrink model size and memory traffic, lowering latency without harming accuracy. Finally, a response-customization module dynamically refines system outputs based on the recognized emotions. The experimental implementation on actual edge devices demonstrates that the proposed system effectively adapts the interaction types to the emotions of the users. In addition, the proposed model achieves low latency and resource-efficient performance.

1. Introduction

VOICE-BASED user interface (VUI), also known as a voice agent, is a verbal and auditory interaction with a computer-based system with voice as input and output. Several researchers are making continuous efforts to develop VUI and its related voice-based technologies. Such technological advancements aim to have a seemingly more natural communication with VUIs as if a human is socially interacting with another person [1–3]. To achieve that, the VUI systems must understand the context of the users. VUI systems without an understanding of the user context often cause issues in performing tasks and usually fail to meet the expectations of humans. To resolve this issue, some researchers proposed an adaptive VUI that automatically adjusts the system parameters based on the understanding of the user context [4–6].

Previous studies have sought to establish new ergonomic criteria for VUIs by modeling interpersonal conversations [7]. They also emphasize the need to account for individual differences in human characteristics [8]. In practice, people prefer different interaction styles; thus, user characteristics and preferences must be incorporated for optimal interaction. Beyond traits and preferences, accurately inferring the user's emotional state is also crucial for configuring VUI parameters and responses. Emotions are mental states associated with internal chemical and physiological changes and with external expressions such as facial expressions, posture, movements, and speech [9]. Although emotions are not directly measurable, they can be inferred from expressive cues, self-reports, physiological indicators, and context. Numerous methods leverage both physiological and non-physiological signals [10]. Non-physiological signals including speech, facial expression, gesture, and pose [1,10–16] are easy to capture but can yield biased estimates [17,18]. Consequently, many works advocate the use of physiological signals [19–22], such as electroencephalograms (EEG), electrooculograms, electrocardiograms (ECG), galvanic skin response,muscle activity or electromyogram, photoplethysmography (PPG), pulse of blood volume and respiration. Nevertheless, relying on a single signal type still limits recognition accuracy.

* Corresponding author.
E-mail address: wansu.lim@skku.edu (W. Lim).

https://doi.org/10.1016/j.knosys.2026.116458
Received 5 December 2025; Received in revised form 14 May 2026; Accepted 13 June 2026
Available online 25 June 2026

To address the limitations of single modalities, multimodal emotion recognition has been widely explored. Chang et al. [23] combined facial expression, skin conductance, finger temperature, and heart rate to achieve 95% accuracy. Likewise, Dutta and Ganapathy [24] proposed a dual-encoder that fuses semantic and acoustic representations for speech emotion recognition, improving generalization across diverse datasets. Moreover, multimodal approaches have been shown to outperform single modalities such as facial expression combined with either PPG or EEG alone [21,25]. Accordingly, in these studies, emotions elicited during VUI interactions are recognized by pairing EEG or PPG with facial expressions, enabling deeper access to underlying affective states. By processing physiological and non-physiological cues together, advanced interfaces can infer user emotions with greater fidelity. Prior VUI-focused efforts illustrate this direction: Hu et al. [26] used acoustic speech features to classify sentiment (positive/negative/neutral) during interactions with conversational agents and to generate empathetic responses. Swoboda et al. [27] compared physiological features (e.g., electrodermal activity, facial micro-expressions) and speech features to predict the intensity of users' emotional responses.

Despite these advances, several gaps remain. Although multimodal signals have been shown to improve emotion recognition, many existing approaches primarily focus on recognition accuracy rather than on how the recognized emotional state can be translated into adaptive VUI responses under edge-device constraints. Non-physiological cues are vulnerable to bias and subjectivity [17,18]. Fusion is often shallow or late, underutilizing complementarities between EEG/PPG and facial features [21,25]. Personalization remains limited despite clear inter-individual differences [7,8]. Moreover, on-device constraints are frequently overlooked, limiting the practical deployment of emotion-aware VUI systems [26,27]. These observations suggest that the representation of emotion is decisive for VUI performance. Therefore, we move from discrete labels to a dimensional formulation to enable continuous tracking. This shift supports edge-suitable policy design and motivates pairing EEG with facial features.

Aside from improving the performance of VUI by using multimodal signals, different types of emotions can also be considered. That is, the accuracy of emotion recognition also depends on the type of emotions that are recognized. Most studies in literature classify emotions in a discrete or dimensional way. Discrete emotions include happiness, sadness, surprise, fear, disgust, and anger. On the other hand, dimensional emotions include dimensions such as valence, arousal, dominance, and liking. Using discrete emotions is less complex than using dimensional emotions. However, dimensional emotions offer a more fundamental and detailed explanation of the relationship between emotions and behaviors. Among dimensional emotions, arousal and valence are commonly used in the field of psychology. In fact, several researchers have classified emotions using the dimensions of valence, arousal, dominance, and liking [28–30]. Subsequently, each discrete emotion is plotted on a 2D plane, with arousal and valence serving as the horizontal and vertical axes, respectively.

The types of emotions recognized by the VUI system directly affect its performance. Specifically, basic discrete emotions such as fear and disgust are rarely observed in short, query-oriented interactions. Additionally, observing the intensity and temporal variations of emotions during the interaction is important for understanding the overall emotional state. Given these challenges, our approach shifts from discrete labels to a dimensional representation, enabling a more fine-grained extraction of emotional output. While arousal and valence can capture continuous changes in user emotion, assessing user satisfaction using only traditional VUI methods has limitations. To overcome this, we introduce two user-centric dimensions: stability and favorability to quantify satisfaction. To effectively measure these dimensions, we integrate facial video and EEG data. Within this context, EEG is adopted as the physiological modality of choice. It is non-invasive, cost-effective, and offers millisecond-level temporal resolution for rapid emotional changes. In contrast to fMRI and fNIRS, which exhibit high latency and limited portability, EEG is available via wearable devices and integrates readily with edge-based VUI systems [31,32]. In addition, peripheral biosignals such as ECG and PPG reflect slower autonomic dynamics and are more susceptible to contextual confounds. By contrast, EEG offers more direct neural correlates of arousal and valence for short, query-oriented interactions. Facial video captures subtle expressions that indicate favorability, while EEG provides insights into the stability of the user's emotional state [32]. By combining these modalities, the system estimates both dynamic emotional changes and overall satisfaction via the predicted 'favorability' and 'stability' scores.

Regarding the implementation of on-device VUIs, several studies have investigated edge-based approaches [33–36]. However, a complete end-to-end deployment on edge devices has not yet been demonstrated. In this work, we present a comprehensive process from signal generation to user-adaptive response optimized for efficient edge implementation. In this context, computational efficiency becomes a key requirement for practical edge-based VUI deployment. To reduce the computational complexity of the proposed multimodal VUI system, a state-of-the-art pruning scheme is integrated into the framework. In particular, cross-layer ranking with k-reciprocal nearest filters (CLR-RNF) [37] is employed. Unlike conventional magnitude- or heuristic-based pruning approaches [38–40], the relative importance of filters across layers is evaluated using a cross-layer ranking strategy in CLR-RNF. Redundant filters are subsequently eliminated via a k-reciprocal nearest selection mechanism. This process effectively preserves the representational diversity of convolutional filters while substantially reducing the overall parameter count and floating-point operations (FLOPs). Combined with post-training quantization techniques [40–42], the pruning process enables the proposed VUI model to maintain high emotion recognition accuracy under strict latency and memory constraints. Consequently, the final network achieves a practical balance between recognition performance and computational efficiency, making it suitable for multimodal inference on resource-constrained edge platforms.

The above studies provide important foundations for emotion-aware VUI systems, multimodal emotion recognition, and efficient model deployment. However, most existing approaches address these aspects separately. Studies on adaptive VUIs mainly emphasize user-context-aware interaction, but they often lack a complete pipeline that links multimodal emotion inference to concrete response adaptation. Multimodal emotion recognition methods improve the robustness of affective state estimation by combining physiological and non-physiological cues, but many of them focus primarily on recognition accuracy without fully considering real-time edge constraints. Meanwhile, pruning and quantization techniques effectively reduce model size and computational cost, but they are usually investigated as standalone model-compression methods rather than as part of an emotion-aware VUI system. Therefore, a unified framework that jointly considers multimodal emotion representation, adaptive VUI response generation, and edge-oriented computational efficiency is still needed.

Building on the motivations and limitations mentioned above, this paper develops a compact, emotion-aware VUI system optimized for edge devices. The primary contributions are as follows:

- **Lightweight multimodal emotion-aware VUI framework for edge devices.** This paper presents a compact framework that integrates physiological (EEG) and non-physiological (facial expression) signals. Our response-customization module provides the adaptive interactions based on recognized emotions.
- **Edge-oriented model optimization with CLR-RNF pruning and post-training quantization.** Per-layer sparsity is determined via a computation-aware importance metric and cross-layer ranking. The retained filters are selected by the k-reciprocal nearest-filter rule. After pruning, uniform INT8 quantization is applied, yielding 4× lower memory and reduced latency.

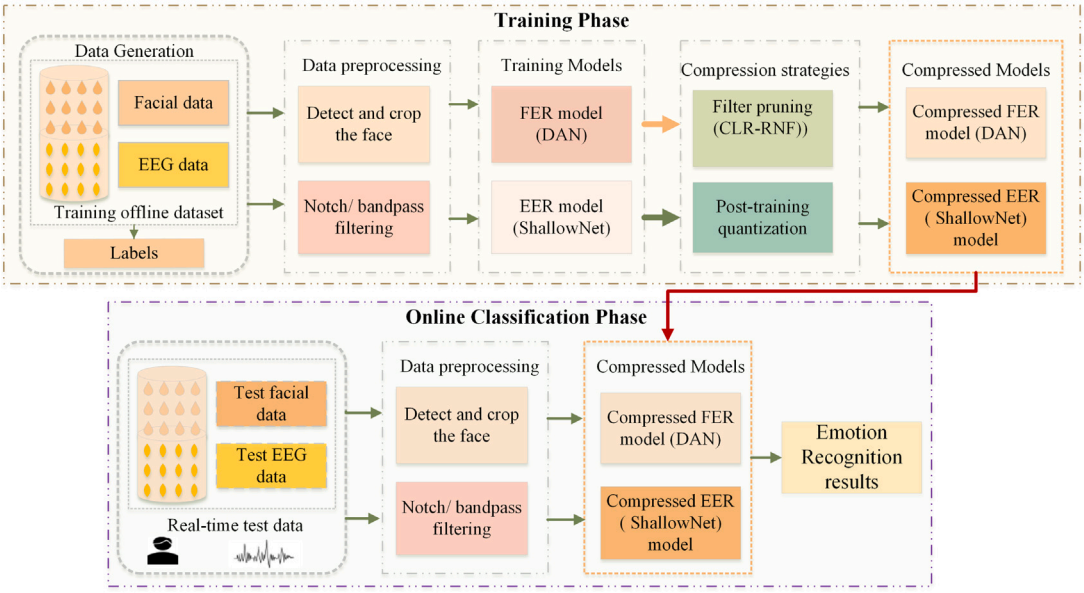


Fig. 1. Block diagram of developing emotion recognition models with two modalities, facial expression and EEG signals.

- **Two novel dimensional emotion axes: stability and favorability.** These axes provide a continuous, interpretable representation that maps to user satisfaction levels. They capture emotional coherence and evaluative valence that conventional discrete-category models do not express effectively.

2. Emotion recognition framework

The paper proposes a user-adaptive VUI system capable of automatically adjusting its interaction style based on the user's emotional state. A multimodal emotion recognition model using facial expressions and EEG is developed, consisting of training and classification phases. Emotions are represented along two axes: stability and favorability. Which are mapped to satisfaction levels to determine whether the user approves of the VUI's response type. Multimodal signals and emotion variations from different interaction styles are used to train a neural network to classify satisfaction states, while a scheduling mechanism extracts temporal features for synchronized adaptation. To support deployment on resource-limited edge devices, the system integrates model optimization techniques such as CLR-RNF pruning and INT8 post-training quantization, enabling efficient real-time inference with minimal performance loss. The lightweight framework is designed with edge constraints in mind, forming a basis for future emotion-aware VUI applications. The overall framework is summarized in Fig. 1.

2.1. Model development

This section presents the development of emotion recognition model that infers the user's emotion based on multimodal signals, consisting of facial expression and EEG signals. In the training phase, the facial expression-based emotion recognition (FER) and EEG-based emotion recognition (EER) models are developed by training each neural network using the multimodal training data labeled with emotions. Each collected data is preprocessed by detecting and cropping face or filtering EEG data before feeding it to the neural network. For the neural network, distract your attention network (DAN) [43] and ShallowNet [44] architectures are employed to build the FER and EER model, respectively. This architectural choice is driven by the need to achieve a balance between recognition performance and computational efficiency, which is critical for real-time edge deployment. Although transformer-based models have recently demonstrated strong performance in modeling long-range dependencies, they are typically associated with high computational complexity, large memory footprint, and significant energy consumption. Such characteristics pose challenges for deployment on resource-constrained edge systems without additional optimization [45,46]. Furthermore, in the context of EEG analysis, transformer-based models often require large-scale datasets to ensure stable training and may suffer from overfitting and unstable generalization under data-constrained and noisy conditions, which are common characteristics of EEG signals [47]. Therefore, this work employs lightweight and modality-specific architectures to ensure efficient and reliable on-device inference. In the classification phase, facial expression and EEG data are acquired in real-time and preprocessed to be fed into the FER and EER models. Each model is deployed in the VUI system in order for emotions to be recognized by the models. In this work, two primary emotion axes are used: stability and favorability. The favorability is inferred based on the facial expression and the stability is inferred based on the EEG signal. Each emotion is recognized by obtaining satisfaction states consisting of satisfied and dissatisfied. Depending on the satisfaction state, stability can be classified as either 'stable' or 'unstable' and favorability can be classified as either 'favored' or 'unfavored'. The detailed classification methods are discussed in the following.

Fig. 2 illustrates the DAN architecture used in our FER model. DAN is a multi-head cross-attention network comprising a feature extraction phase and an attention phase. The feature extraction phase employs a feature clustering network (FCN) built on a ResNet18 backbone [48] pretrained on the MS-Celeb-1M dataset [49], and is supervised through contrastive learning to obtain a more discriminative 512-dimensional feature representation. The attention phase includes two modules: the multi-head cross attention network (MAN) and the attention fusion network (AFN). MAN contains parallel spatial and channel attention units (blue and green in Fig. 2), using convolutional layers for spatial features and a fully connected layer for channel features to generate multiple attention maps. AFN then merges these maps to avoid redundancy (red in Fig. 2) and produces the final prediction feature map, which is passed to a dense layer for classification. The DAN outputs class probabilities via SoftMax. Further details on the DAN framework can be found in [43].

To obtain feature representations better suited for our study, the DAN architecture replaces its original ResNet18 backbone with

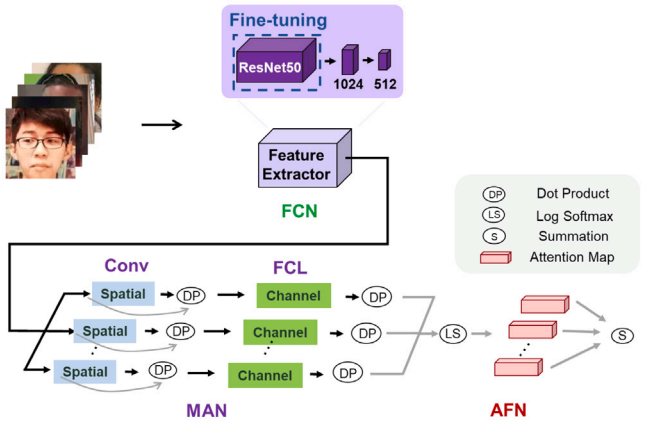


Fig. 2. An architecture of DAN utilized for the facial-expression based emotion recognition model.

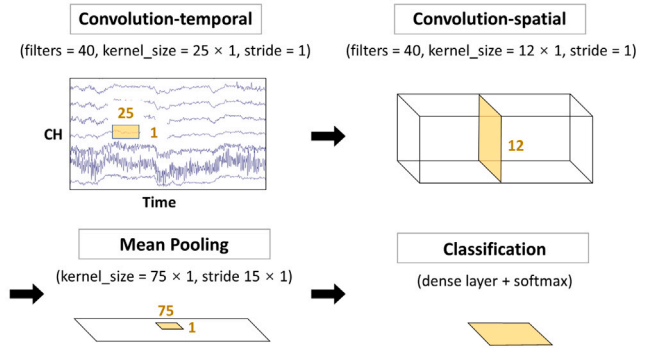


Fig. 3. Architecture of ShallowNet utilized for the EEG based emotion recognition model.

ResNet50 in the FCN module, as shown in Fig. 2. ResNet50 is pretrained on the large-scale VGGFace2 dataset, and its effectiveness for in-the-wild emotion classification was demonstrated in our previous work [50,51]. Using DAN with ResNet50, we achieved competitive results on the challenging 8-class facial expression recognition task defined in [52] (six basic emotions plus "neutral" and "other"), outperforming a baseline method [53] (VGG16 [54] pretrained on VGGFace [55]) by 11.6% mean F1-score and ranking second in the challenge [53]. Building on this improvement, DAN with ResNet50 is fine-tuned on our custom dataset for binary satisfaction classification. The model is retrained using a batch size of 120, the Adam optimizer (learning rate 1e−4), and 40 epochs on a high-end multi-GPU server. The resulting FER model has 45,962,448 trainable parameters, which are kept frozen in the final study.

For the EER model, a shallow convolutional neural network (ShallowNet) is adopted, consisting of two convolutional layers, one pooling layer, and one dense classification layer, as shown in Fig. 3 [44]. Designed to mimic the widely used FBCSP method, ShallowNet extracts spatial–spectral–temporal EEG features: the first and second convolutional layers act as temporal and spatial filters, while the subsequent layers perform mean pooling and classification. ReLU is used as the activation function for all layers, and the final fully connected layer outputs class probabilities via SoftMax. The EEG dataset is divided into training and validation sets (8:2), and the network is trained with the Adam optimizer (learning rate 0.001) for 200 epochs, minimizing MSE loss. Dropout (0.5) and batch normalization (0.1) are also applied. The final EER model contains 21,922 trainable parameters, which are kept frozen in this study.

2.2. Filter pruning and quantization

To enable lightweight deployment on resource-constrained edge platforms, this paper adopts a two-stage pipeline: structured filter pruning via CLR-RNF, followed by post-training quantization, applied to both FER and EER models. The pruning stage begins with a computation-aware importance measure for individual weights to avoid magnitude-only bias. Next, cross-layer ranking is performed to identify and remove bottom-ranked weights, which induces per-layer sparsity that defines the pruned structure. On top of this structure, a recommendation-based filter selection is used: each filter recommends its nearest peers. Preserved filters are then chosen by the k-reciprocal nearest-filter rule, i.e., the intersection of mutual-neighbor groups. Both the structure derivation and k-reciprocal nearest-filter selection are non-learning procedures, substantially reducing pruning complexity and ensuring consistent cross-layer sparsity while preserving salient representations. After pruning, post-training quantization is applied to enable 8-bit inference, reduce memory footprint (4×), and lower latency without retraining. The setup uses uniform 8-bit, per-channel activations, per-tensor weights, and a calibration set to estimate scale. The resulting compressed models are denoted Q-FER and Q-EER. In this study, the FER network with 45,962,448 parameters undergoes a global pruning ratio of about 60%. The EER network with 21,922 parameters is already lightweight. Therefore, only 10% minor pruning is applied to preserve model stability. After pruning, both networks are uniformly quantized to 8-bit precision per-channel for activations and per-tensor for weights, resulting in a nearly fourfold reduction in memory footprint. This combined pruning–quantization strategy yields compact FER and EER variants that preserve accuracy while achieving low-latency, memory-efficient inference suitable for edge device deployment.

2.2.1. Structured filter pruning

Let the pretrained FER and EER networks contain L convolutional layers, $C = \{C_1, C_2, \dots, C_L\}$. Each layer C_i includes n_i filters $\mathbf{W}_i = \{\mathbf{w}_i^1, \mathbf{w}_i^2, \dots, \mathbf{w}_i^{n_i}\}$, where $\mathbf{w}_i^j \in \mathbb{R}^{c_{i-1}\times h_i \times w_i}$ denotes the jth filter in layer i. The pruning goal is to produce a compact model $\bar{\mathcal{W}} = \{\bar{\mathbf{W}}_1, \dots, \bar{\mathbf{W}}_L\}$, where each $\bar{\mathbf{W}}_i$ retains $\bar{n}_i = \lfloor (1-p_i)n_i \rfloor$ filters according to the adaptive per-layer pruning rate p_i.

(a) Computation-aware importance: The importance of each weight element is defined as

$$\alpha_{i,q}^j = \frac{|(\mathbf{w}_i^j)_q|}{(\mathrm{FLOPs}_i)^\lambda}, \tag{1}$$

where FLOPs_i denotes the computational cost of layer i and $\lambda \geq 0$ balances sparsity and efficiency. Smaller $\alpha_{i,q}^j$ indicates less important weights.

(b) Global pruning and layer-wise sparsity: Weights with the lowest $\alpha_{i,q}^j$ are removed according to the global pruning rate p:

$$(\bar{\mathbf{w}}_i^j)_q = \begin{cases} 0, & \text{if } \alpha_{i,q}^j \text{ belongs to the lowest-ranked set,} \\ (\mathbf{w}_i^j)_q, & \text{otherwise.} \end{cases} \tag{2}$$

The effective pruning ratio of layer i is

$$p_i = 1 - \frac{\sum_{j=1}^{n_i}\sum_{q=1}^{c_{i-1}h_i w_i} \delta((\bar{\mathbf{w}}_i^j)_q \neq 0)}{n_i\, c_{i-1} h_i w_i}, \tag{3}$$

where $\delta(\cdot)$ equals 1 when the condition holds and 0 otherwise.

(c) Reciprocal filter selection: To preserve feature diversity, a *k-reciprocal nearest filter* scheme is adopted. The similarity between two filters $\mathbf{w}_i^j$ and $\mathbf{w}_i^h$ is given by

$$S_i^{jh} = \frac{\exp(-\|\mathbf{w}_i^j - \mathbf{w}_i^h\|_2^2)}{\sum_{g=1}^{n_i}\exp(-\|\mathbf{w}_i^j - \mathbf{w}_i^g\|_2^2)}. \tag{4}$$

The k nearest neighbors of each filter are defined as

$$\mathcal{N}^k_{\mathbf{w}_i^j} = \{\mathbf{w}_i^h \mid R_i^{jh} \le k\}, \quad R_i^{jh} = 1 + \sum_{g=1}^{n_i} \delta(S_i^{jg} > S_i^{jh}). \tag{5}$$

The pruned filter set of layer i is obtained as

$$\bar{\mathbf{W}}_i = \bigcap_{j=1}^{n_i} \mathcal{N}^k_{\mathbf{w}_i^j}, \tag{6}$$

ensuring that only mutually close filters in feature space are retained. In practice, the parameter k is not fixed but adaptively determined for each layer based on the target number of retained filters $\bar{n}_i$. Specifically, k is progressively increased until the cardinality of the selected filter set satisfies $|\bar{\mathbf{W}}_i| = \bar{n}_i$. This adaptive strategy accounts for layer-wise variations in filter similarity distributions and avoids the limitations of a fixed k, which may lead to over-pruning or redundant filter retention. Finally, the pruned network is fine-tuned to recover potential accuracy loss.

2.2.2. Post-training quantization

After pruning, all convolutional weights are uniformly quantized to 8-bit integers. For a full-precision weight w_f, quantization and dequantization are defined as

$$w_q = \text{round}\left(\frac{w_f}{s}\right), \quad \hat{w}_f = s \cdot w_q, \tag{7}$$

where $s = \frac{\max(|w_f|)}{2^{b-1}-1}$ is the scaling factor for bit-width $b = 8$. Per-channel dynamic ranges are estimated for activations using a small calibration set.

The complete pipeline is summarized in Algorithm 1. The resulting Q-FER and Q-EER models achieve significant reductions in parameters and FLOPs, enabling real-time multimodal inference on edge devices with minimal accuracy loss.

This pruning-quantization design reflects the trade-off between model complexity and computational efficiency in the proposed edge-oriented VUI system. A higher pruning ratio can further reduce parameters and FLOPs, but excessive pruning may remove salient filters and degrade recognition accuracy. Therefore, different pruning ratios are assigned according to the computational characteristics of each branch: the FER branch is compressed more strongly as the main computational bottleneck, whereas the EER branch is only lightly pruned to preserve model stability. Regarding scalability, the compressed Q-FER and Q-EER models are shared across VUI interaction tasks, so increasing the number or diversity of tasks does not necessarily require separate emotion recognition backbones. Instead, task diversity mainly affects the response adaptation stage, where recognized emotional states are mapped to a larger set of response types or interaction strategies. If future applications require additional modalities, finer-grained emotion categories, or more complex task-specific reasoning modules, the same CLR-RNF pruning and INT8 quantization strategy can be extended to newly added modules to maintain real-time and memory-efficient edge operation.

In this study, each emotion recognition model classifies the user's satisfaction state, either satisfied or dissatisfied, using a SoftMax output layer. SoftMax normalizes the outputs to probabilities between 0 and 1, and the class is determined by a 0.5 threshold: probabilities above 0.5 are labeled as satisfied, otherwise dissatisfied. These satisfaction states are then mapped onto the two-dimensional emotion plane (Fig. 4). For example, if both stability and favorability are classified as satisfied, the emotion is mapped to the "stable–favored" quadrant. Two modalities are used: facial expression data for favorability (handled by the compressed FER model) and EEG signals for stability (handled by the compressed EER model). The resulting probability scores are placed on the x- and y-axes of the emotion plane. Performance of both models is evaluated through 5-fold cross-validation on a subject-specific offline dataset.

Algorithm 1 Computation-aware filter pruning and quantization

Input: Trained FER and EER networks $\mathcal{W} = \{\mathbf{W}_1, \dots, \mathbf{W}_L\}$, global pruning rate p
Output: Compressed and quantized models (Q-FER, Q-EER) for real-time inference

Stage 1: Structured filter pruning
Compute weight importance
for $i = 1$ to L **do**
  for $j = 1$ to n_i **do**
    for each weight $(\mathbf{w}_i^j)_q \in \mathbf{w}_i^j$ **do**
      Compute importance $\alpha_{i,q}^j = \frac{|(\mathbf{w}_i^j)_q|}{(\text{FLOPs}_i)^\lambda}$
    end for
  end for
end for
Global ranking and pruning
Rank all weights by $\alpha_{i,q}^j$ and prune lowest p% via Eq. (2)
Compute per-layer pruning rate p_i using Eq. (3)
kk-reciprocal nearest filter selection
for $i = 1$ to L **do**
  Initialize $k = \lfloor (1 - p_i) n_i \rfloor$, $\bar{\mathbf{W}}_i = \emptyset$
  while $|\bar{\mathbf{W}}_i| \neq \bar{n}_i$ **do**
    Compute similarity S_i^{jh} using Eq. (4)
    Compute closeness rank R_i^{jh} and
    Form neighbor sets $\mathcal{N}^k_{\mathbf{w}_i^j}$ using Eq. (5)
    Obtain intersection $\bar{\mathbf{W}}_i = \mathcal{N}^k_{\mathbf{w}_i^1} \cap \cdots \cap \mathcal{N}^k_{\mathbf{w}_i^{n_i}}$
    $k \leftarrow k + 1$
  end while
end for
Stage 2: Post-training quantization
for each weight tensor $\bar{\mathbf{W}}_i$ **do**
  for each element $w_f \in \bar{\mathbf{W}}_i$ **do**
    Compute scale factor $s_i = \frac{\max(|w_f|)}{2^{b-1}-1}$ with $b = 8$
    Quantize $w_q = \text{round}(w_f / s_i)$
    Dequantize $\hat{w}_f = s_i \cdot w_q$
  end for
end for
Fine-tune the quantized model to recover accuracy
return $\hat{\mathcal{W}}$ (Q-FER and Q-EER)

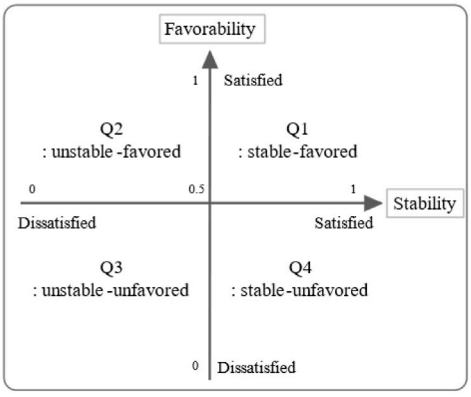


Fig. 4. Emotion plane where the models classify the emotion into one of four quadrants (Q1, Q2, Q3, Q4) and map the probability value of satisfaction.

2.3. Data preprocessing

To construct the emotion recognition models, facial images and EEG signals are processed using separate pipelines. Facial data are first trimmed to remove background noise, and DeepFace [56,57] is used to

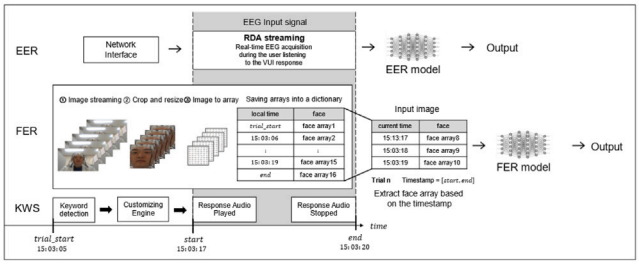


Fig. 5. Scheduling mechanism in proposed framework.

detect and crop the face region. Each cropped face is then resized to 224 × 224 before being fed into the FER model. EEG preprocessing is performed using SciPy and MNE-Python [58,59]: raw data are segmented, notch-filtered at 59–61 Hz, bandpass-filtered between 1–80 Hz, and downsampled from 1,000 Hz to 200 Hz to reduce computational cost. Fourteen channels located around the frontal, temporal, and temporal-central regions (FP1, FP2, FT9, T7, TP9, TP10, T8, FT10, FT7, C5, TP7, TP8, C6, FT8) are used for model training. These channels are selected from the sixty-four collected channels to reduce computational cost while preserving EEG regions relevant to emotion recognition [60], following prior reduced-channel EEG emotion-recognition studies [61]. Artifacts caused by noise and user movement are removed using ICA implemented in MNE-Python [59].

2.4. Task scheduling

In this paper, scheduling is employed to extract the data from different modalities at a certain timing. The input data to feed the emotion recognition model is a bimodal signal that is collected while the user listens to the VUI's response. For this purpose, the proposed VUI requires a scheduling mechanism to extract data by capturing the moment of the start and end of the response. Fig. 5 shows the process of extracting each signal collected in valid timing, i.e., during the response audio playback. In terms of facial input data, the timestamps of the start and end of the VUI response audio playback are saved as the local variables. Simultaneously, in the FER code block, every image frame streamed per second is stored as an array in a key–value pair format along with the current timestep. The incoming image streaming data are trimmed to crop the face. Thereafter, this cropped face image is resized and converted into an array and stored with the timestamp in a key–value pair format. By doing so, facial data collected during the audio playback is extracted based on the time information. As an example, if the trial start time, audio start time, and audio end time are given as 15:03:05, 15:03:17, and 15:03:20, respectively, the image frame from 15:03:17 until 15:03:20 are extracted only. Therefore, considering that only one frame is selected per second, only three frames are extracted and supplied to the FER model for classification. On the other hand, valid EEG input data is extracted by monitoring the audio playback using the pygame library, which can start the playback of the audio file and monitor whether the audio is playing or not. While the EEG recording device records the EEG signals, real-time data is simultaneously transmitted from the recording device to the edge device, where a VUI system is deployed. To this end, remote data access (RDA) program, which can be activated in the recording software, is utilized to stream EEG signals from recording device to the external device located in the local network area. Therefore, the network is placed between the EEG recording device and the VUI system to acquire the EEG signals in real-time before starting the interaction. With the audio monitoring library, the response audio of the VUI is played and streaming of the EEG signals is activated only while the audio is playing using RDA program. Finally, EER and FER models can successfully extract valid input data.

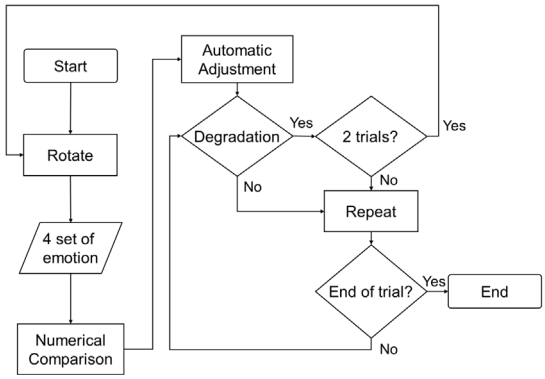


Fig. 6. Flow chart of proposed customizing engine.

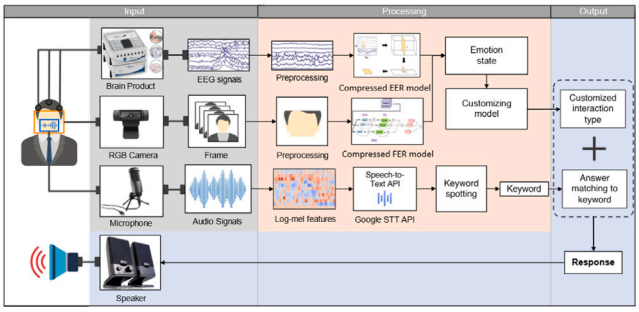


Fig. 7. Scheduling mechanism in proposed framework.

2.5. Customizing model

The customizing model within the VUI system provides personalized responses based on the user's emotions, which are evaluated at the end of each trial using the emotion recognition models. Emotions are mapped onto a two-dimensional plane defined by stability and favorability, with the origin at (0.5, 0.5), representing the satisfaction probabilities of each axis. An emotion is assigned to one of four quadrants, with the first quadrant indicating satisfied states for both axes (probabilities higher than 0.5). The goal of the customizing model is to identify the response type that consistently leads to the first-quadrant outcome. As shown in Fig. 6, the model rotates through four response types in a random order over four trials, collects the corresponding emotion results, and selects the response type with the highest satisfaction probability. Once the optimal type is found, the system continues using it in subsequent trials until a decline in satisfaction is detected, ensuring dynamic and personalized interaction.

3. VUI system implementation

The overall system configuration of the proposed user-adaptive VUI is visualized in Fig. 7. The figure shows the hardware and software components of the input, processing, and output blocks. After acquiring input data, each data is subjected to the pre-processing techniques in the processing block, as illustrated in Fig. 7. These preprocessing methods are discussed in detail below in Section 3.1. The preprocessed input data is utilized by the FER and EER models for the recognition of the user's emotional state. The predicted emotion state is fed into the customizing model, which identifies a personalized interaction type. The proposed VUI incorporates the functionality of keyword. When the user verbally asks a question to VUI, the system recognizes the speech. To this end, the speech features are analyzed and then transcribed into text. Finally, the VUI generates a response by combining the customized response type and spotted keyword. The response is provided to the

Table 1
Database content summary.

Subject annotation	
Parameter of gender in response	Male, female
Parameter of information quantity in response	Reference, short
Gender of participants	Male, female
Rating scales/values	6 adjective pairs, scale −3 to 3
Emotion score	Stability, Favorability
No. of the trial	120
Experiment details	
Number of subjects	35
Number of trials	120
Recorded signals	63-ch 1000 Hz EEG; Face video

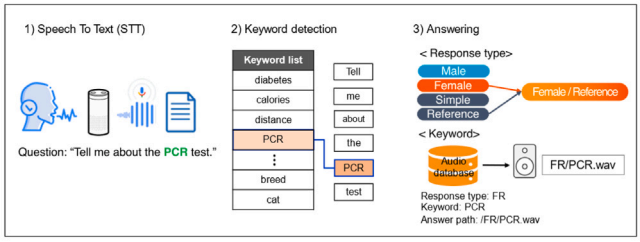


Fig. 8. Keyword spotting-based answer retrieving mechanism of the proposed VUI system.

user through the connected speaker, as shown in the output block. The process described above enables users to have iterative interactions with the VUI. The details are presented in the following subsections.

3.1. System configuration

To achieve portability and real-time performance, most processes are deployed on an embedded edge device equipped with a high-performance GPU suitable for AI and parallel processing. The system acquires data in real time from two modalities: facial expressions and EEG signals. For facial data, a web camera (2 megapixels, 1920 × 1080 resolution) operating at 30 fps is used, and one frame per second is selected as input to the FER model. EEG signals are recorded using a standard EEG amplifier, with conductive gel applied at each electrode site prior to measurement. After electrode placement, scalp-electrode impedance is checked, and recording begins only when the impedance is below 25 kΩ. Fourteen active electrodes are used, and the EEG signals are sampled at 1000 Hz, resulting in a continuous data stream of 1000 × 14 samples per second.

3.2. Keyword spotting

The proposed VUI supports query-based interaction, where the user asks a question and the system responds using one of four interaction types: male/simple (MS), male/reference (MR), female/short (FS), and female/reference (FR). To ensure natural communication, the user's speech is processed in real time. As shown in Fig. 8, the answer-searching mechanism includes three steps: (1) speech-to-text conversion, (2) keyword detection, and (3) response generation. After the system transcribes the spoken input into text, the sentence is segmented into words and matched against a predefined keyword list. Based on the detected keyword, the system selects both the appropriate response and the corresponding interaction type.

3.3. Generation of training dataset

For the training database, multimodal signals are collected while the participants' emotions are triggered by the distinct types of response.

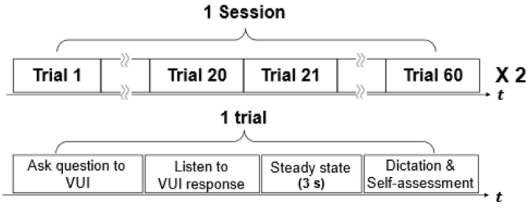


Fig. 9. Schematic of the number of trials per session and detail procedure of each trial.

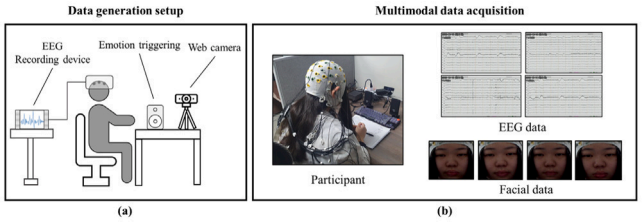


Fig. 10. Presentation for multimodal data acquisition of EEG and facial data while the participant is emotionally triggered by different design parameters: (a) data generation setup; (b) actual raw data acquired from two modalities, including EEG and facial data.

To this end, informative questions were randomly selected by the participants to ask to the virtual prototype which mimics VUI interactions. The database contains all recorded EEG signal data and frontal face video for a subset of the participants. Also included is the subjective ratings of provided response type which varies in its parameters of gender and information quantity. Furthermore, statistical analysis of the participant's ratings is presented via considering correlations between the ratings of emotional adjectives. By doing so, score for each emotion axis of stability and favorability are obtained and annotated with the multimodal signal for each trial. Table 1 gives an overview of the database contents.

3.3.1. Overall procedures

Thirty-five participants (18 males, 17 females), all Korean students from Kumoh National Institute of Technology with an average age of 22.6, took part in the experiment. While the participants were recruited from a relatively homogeneous demographic group, this controlled setting was intentionally adopted to ensure consistency in multimodal data acquisition (EEG and facial signals), which can be sensitive to environmental and subject variability. As summarized in Fig. 9, the study consisted of two sessions on different days, each containing 60 trials. A moderator guided participants throughout the process. In each trial, the participant asked one question selected from a pool of 240, after which the VUI detected keywords and provided a response within about 1 s using specific design parameters (voice gender and response length). Following the response, a 3-second pause allowed physiological signals to stabilize. At the end of each trial, participants repeated the response they heard and rated their satisfaction through a self-assessment survey. These ratings were converted into emotion scores for the two emotion axes and used as ground-truth labels for the corresponding emotional states. Although the current dataset is limited in demographic diversity, the proposed framework is not restricted to this specific population and can be extended to more diverse user groups in future studies.

3.3.2. Emotion triggering using design parameters

The multimodal dataset is gathered while the subjects are interacting with the VUI system. The dataset generation procedure aims to determine the emotional responses triggered by different response

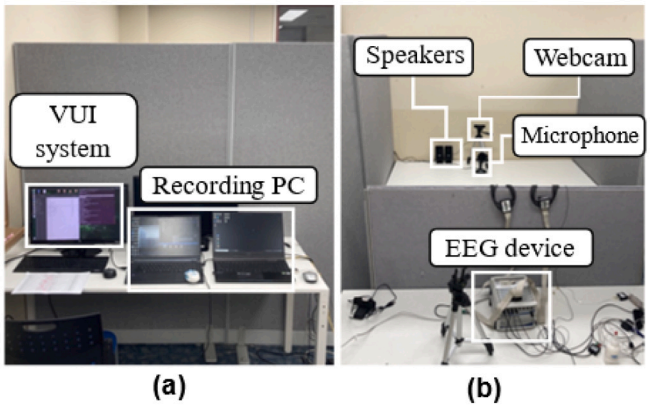


Fig. 11. The experimental environment consisting of (a): Monitoring space, (b): Interaction space.

types. To this end, the VUI intentionally changes the response type by giving variations in several parameters of the voice. Inspired by prior work on VUI design parameters [62], six voice design parameters are initially considered in this study: gender (male, female), age (child, adult), honorific style (formal, informal), pitch (high, mild), information quantity (simple, reference), and confidence (confident, timid). Among all design parameters, it is determined that the variations in 'gender' and 'information quantity' showed significant effects on the physiological signal patterns. Hence, the VUI randomly combines those two parameters which are subsequently used to respond to the users. The response type combination includes: male/simple (MS), male/reference (MR), female/short (FS), or female/reference (FR). For gender, adult voices are applied. For response length, a short answer for the simple type and a detailed answer for the reference type are provided. To enable VUI to randomly choose the response type, all answers for every type are prepared in advance.

3.3.3. Multimodal data acquisition

Fig. 9 shows multimodal data acquisition while the emotion of the participant is being triggered by different response types. Fig. 10(a) illustrates the data generation setup, whereas Fig. 10(b) shows the actual sample data acquired from the participant. While the participants are provided with different response types, the face and EEG data are recorded simultaneously. The multimodal data acquisition is conducted in two spaces separated by high partitions: (1) monitoring space and (2) interaction space. Fig. 11 shows the actual environmental setup. In the monitoring space, an embedded computing device and two desktop PCs are placed for recording two different modalities of face and EEG data. In the interaction space, a webcam, EEG amplifier, microphone, questionnaire transcript, and speaker are placed. While the participants interact with the VUI, the moderator monitors the face and EEG data through a webcam and EEG amplifier. The speaker and microphone are connected to the embedded computing device, and the webcam and EEG amplifiers are connected to each dedicated desktop PC. During the whole session, facial and EEG data are continuously recorded. The face data is fully recorded at 30 frames per second. On the other hand, the EEG signal is recorded with a total of sixty-three active electrodes (FP1, Fz, F3, F7, FT9, FC5, FC1, C3, T7, TP9, CP5, CP1, Pz, P3, P7, O1, Oz, O2, P4, P8, TP10, CP6, CP2, Cz, C4, T8, FT10, FC6, FC2, F4, F8, Fp2, AF7, AF3, AFz, F1, F5, FT7, FC3, C1, C5, TP7, CP3, P1, P5, PO7, PO3, POz, PO4, PO8, P6, P2, CPz, CP4, TP8, C6, C2, FC4, FT8, F6, AF8, AF4, and F2). The electrodes are mounted on the participant's scalp based on the international 10–10 system to measure the EEG data. Ground and reference electrodes were attached at Fpz and FCz, respectively.

Table 2
Result of factor analysis for 6 contrasting adjective pairs.

Adjective pairs	Factor 1	Factor 2
Reassured–Concerned	0.810	
Stable–Unstable	0.806	
Certain–Uncertain	0.787	
Comfortable–Inconvenient	0.762	
Impressive–Normal		0.881
Interesting–Uninteresting		0.870
Var	3.8598	2.1883
% Var	0.429	0.243
	Stability	**Favorability**

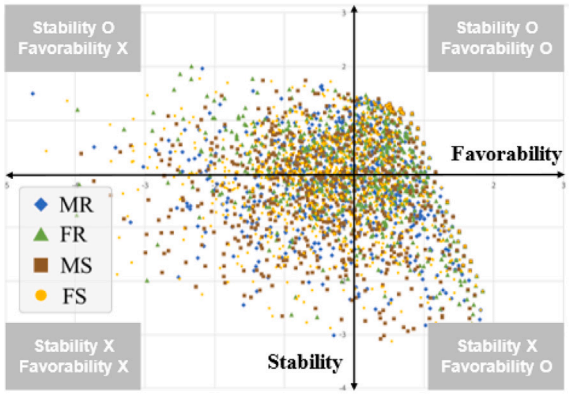


Fig. 12. Classification map of emotion scores on two-dimensional emotion plane with two emotion axes of stability and favorability. Participants' emotions are triggered by four design parameters (MR, FR, MS, FS).

3.3.4. Translation of satisfaction levels

This section explains how the two-dimensional emotion axes: stability and favorability are derived and translated into satisfaction levels. After each interaction, participants evaluate the response type using emotional adjectives, which are analyzed through Kansei engineering [63] to capture emotional impressions. Six key Kansei word (KW) clusters are rated on a 7-point semantic differential scale. Based on correlations among the KW ratings, two underlying psychological factors are extracted using statistical factor analysis, as shown in Table 2, following the procedure described in [64]. The KWs "reassured", "stable", "certain", and "comfortable" load onto factor 1, forming the stability axis, while "impressive" and "interesting" load onto factor 2, forming the favorability axis. It should be noted that the term "factor" here refers to latent psychological dimensions derived from statistical factor analysis, which is distinct from the feature representations learned by the neural networks used in the FER and EER models. Thus, stability and favorability are empirically identified as primary emotional dimensions that collectively represent all six KWs.

Based on these derived emotional axes, ratings of six KWs are converted into factor scores by considering the correlations between the scores of the six adjective pairs. That is, factor scores are obtained as stability score and favorability score to be mapped to the 2-D plane. Fig. 12 shows the mapping of the stability–favorability scores in the 2-D plane for 120 trials of interaction. The four different color-coded points refer to the emotion scores of each trial triggered by the four response types (MR, FR, MS, FS). Considering that each of the 32 participants interacted with the VUI for 120 trials and each design parameter is presented 30 times equally, 3840 sets of stability–favorability scores are mapped in the plane. Based on this map, multimodal signals are classified into four groups according to the four quadrants. To this end, positive emotion scores greater than 0 are considered as satisfied (○), while negative emotion scores less than 0 are classified as dissatisfied (×). Therefore, each set of multimodal signals for one trial is labeled as

Table 3
A complete list of the interaction and corresponding emotional satisfaction.

Trial number	Section of trial	Design parameter	Participant question	Stability	Favorability	Quadrant
1	Rotate	FS	What is the **flower language** of red roses?	0.45	0.35	Q1
2	Rotate	FR	Recommend me good food for **diabetes**.	0.60	0.54	Q1
3	Rotate	MR	How many **calories** are in tomatoes?	0.66	0.32	Q4
4	Rotate	MS	What is the **distance** from the earth to the sun?	0.28	0.35	Q3
5	Adjustment	FR	Where is the **Grand Canyon** located?	0.95	0.72	Q1
6	Repeat	FR	How many **Ikea** stores are in Korea?	0.95	0.86	Q1
7	Repeat	FR	Tell me about the benefits of **red ginseng**.	0.91	0.62	Q1
8	Repeat	FR	What is the **PCR test**?	0.85	0.66	Q1
9	Repeat	FR	Who was the first to coin the term **quantum mechanics**?	0.46	0.39	Q3
10	Degradation	FR	Tell me the number of **public holidays** in Korea this year.	0.46	0.35	Q3
11	Rotate	MR	What materials are needed to make a **mulled wine**?	0.43	0.52	Q2
12	Rotate	FS	Which animal has the longest **lifespan**?	0.36	0.29	Q3
13	Rotate	MS	How many neck bones does a **giraffe** have?	0.46	0.48	Q2
14	Rotate	MR	When is **Joseon National day**?	0.65	0.60	Q1
15	Adjustment	MR	What is the proper **indoor temperature** in winter?	0.96	0.78	Q1
16	Repeat	MR	Tell me about the principle of **spin free kick**.	0.84	0.92	Q1
17	Repeat	MR	What is the official language of the **United Nations**?	0.58	0.95	Q1
18	Repeat	MR	What is the angle of the **Leaning Tower of Pisa**?	0.61	0.59	Q1
19	Repeat	MR	Is the **Dead Sea** a lake or a sea?	0.76	0.64	Q1
20	Repeat	MR	What is the smartest **dog breed**?	0.98	0.88	Q1

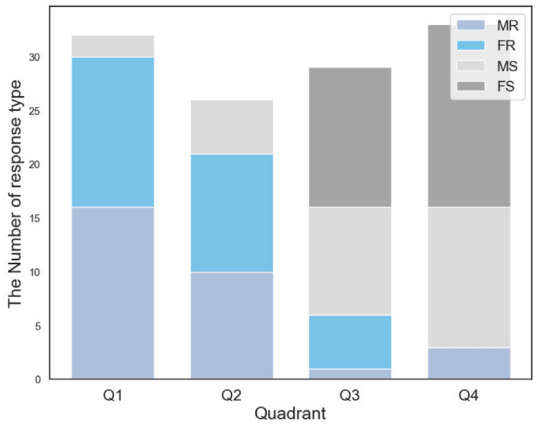


Fig. 13. Frequency of the number of each response type in each quadrant obtained by the subject participated in both procedure of training dataset generation and the demonstration.

Table 4
Comparison of Mean and Std between ground truth and real-time.

Stability				
Type	Mean	Std	Mean	Std
MR	0.639	0.155	0.715	0.213
MS	0.428	0.145	0.405	0.063
FR	0.714	0.158	0.718	0.182
MS	0.425	0.141	0.370	0.127
Favorability				
Type	Mean	Std	Mean	Std
MR	0.694	0.177	0.621	0.148
MS	0.417	0.093	0.435	0.120
FR	0.710	0.209	0.621	0.160
MS	0.412	0.097	0.420	0.084

one out of four labels (S: ○/F: ○, S: ○/F: ×, S: ×/F: ×, S: ×/F: ○) in the stability–favorability plane. By doing so, the classified satisfaction states are translated into final emotion states.

4. Results and discussion

The proposed VUI framework provides user-adaptive feedback by considering the users' emotional and satisfaction state (satisfied/dissatisfied). In the experiment, the user interacted with the VUI system which is designed to trigger a change in the users' emotions. The changes in emotion was measured using the probability of being satisfied or dissatisfied. Table 3 shows the complete list of interaction trials between the VUI and the subject with English-translated user questionnaire speech from Korean. In Table 3, the section of each trial is presented, mentioned above in Section 2.2. In the first trial, the customizing model starts to survey the user's satisfaction rate by randomly picking the response type. In the rotate section, it is shown that the customizing model determines that the 'FR' type is mapped to 'Q1'. This indicates that the corresponding type induced the 'satisfied' state for both emotions of stability and favorability. In the second trial, the emotion recognition result shows the probability higher than the decision boundary 0.5 for both emotion axes. Thus, the customizing model automatically adjusts the response type to 'FR' and repeat the 'FR' type on the following trial. While repeating 'FR', the customizing model simultaneously monitors the probability value and checks if the value degrades more than two trials. In trials 9 and 10, the emotion recognition model captures the degradation of satisfaction for 'FR' type. Therefore, the customizing model re-executes the rotate section during trials 11 to 14. Through this, the model figures out that new type 'MR' induced the highest emotional satisfaction in both stability and favorability. Thus, the customizing model adjusts the response type to 'MR' type and repeats it from trial 15. Based on this testing procedure, it was confirmed that the subject prefers when the VUI interacts with the user in 'FR' and 'MR' type. Fig. 13 summarizes the participant's preferences for response types with a histogram showing the frequency of each response type mapped to each quadrant. Each quadrant is decided based on the stability score and favorability score, which are derived from the participant's self-rated emotion ratings. As shown in Fig. 13, 'MR' type evoked the seemingly high frequency of 'satisfied' state in both emotion axes, such that mapped to the 'Q1' accordingly. Consecutively, 'FR' type showed the second highest frequency in 'Q1'. In contrast to the satisfied emotional responses, 'MS' and 'FS' types incurred relatively less satisfied response from the user. This observation is aligned with the real-time test result, where the proposed VUI provided the 'MR' and 'FR types the most.

For a more thorough comparison, Fig. 14 quantitatively shows the distribution of data of real-time test results addressed above and annotated training dataset. The annotated training dataset contains the subject's self-reported scores for emotion axes defined by the ground truth. Each graph shows the satisfaction level of stability and favorability toward the response types. The emotion outputs predicted by

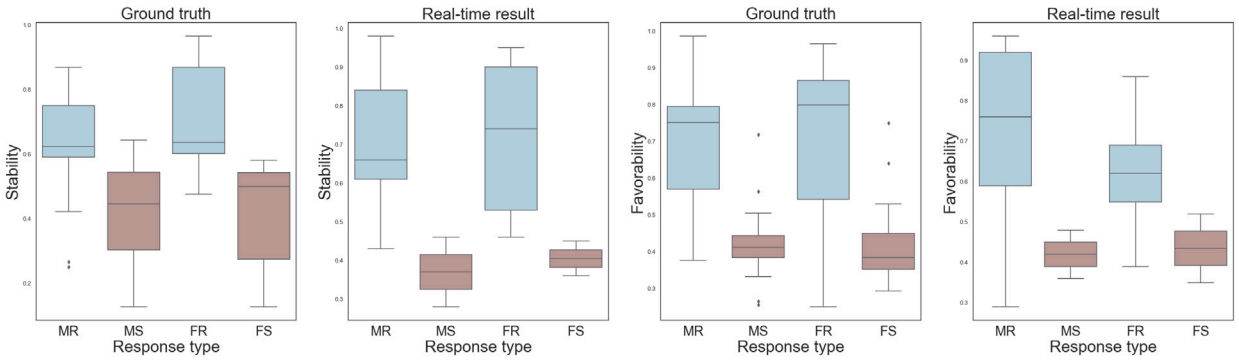


Fig. 14. Comparison of satisfaction level between the real-time test result and the annotated training dataset. The probability from the real-time test and the normalized emotion score are compared for stability and favorability.

the emotion recognition models are presented along with their corresponding response types and are quantified using the probability value listed in Table 3. Table 4 summarizes the comparison of stability and favorability measures for different interaction types under both ground truth and real-time test result conditions. Looking at the mean values of the ground truth and real-time test results, 'MR' and FR' types show relatively higher satisfaction values than the remaining types. In this work, the emotion recognition model produces probabilistic outputs for satisfaction states, which are directly utilized by the adaptive VUI system. Therefore, the evaluation focuses on probability distributions, temporal consistency across interaction trials, and agreement with self-reported ground-truth emotional scores. Unlike conventional standalone classification tasks, the proposed framework is designed as an interactive system, where model outputs are used for real-time response adaptation. While standard classification metrics such as precision, recall, and F1-score can provide complementary insights, they are not the primary evaluation criteria in this work, as they do not fully capture the effectiveness of adaptive interaction behavior.

The experimental results demonstrate that the proposed multimodal VUI framework achieves effective emotion recognition with low latency under on-device deployment. The combination of structured pruning and quantization enables significant model compression while maintaining competitive performance, supporting real-time operation in resource-constrained edge environments. However, several limitations should be noted. First, the current evaluation is conducted on a relatively homogeneous participant group, which may limit generalizability across broader populations. Second, while the system shows stable performance under controlled experimental conditions, its robustness under diverse real-world environments has not been fully evaluated. Factors such as ambient noise, EEG signal variability, and device-level constraints (e.g., thermal throttling and battery limitations) may affect performance in practical deployment scenarios. Basic preprocessing steps (e.g., face detection and EEG filtering) are applied to mitigate moderate noise and signal variations, and the lightweight on-device design supports stable operation under such constraints. Third, the current framework mainly focuses on EEG signals and facial expressions for lightweight emotion recognition, while non-speech auditory cues are not explicitly incorporated. However, non-speech auditory cues can also influence user emotion, attention, and interaction behavior. Previous studies have shown that sonic feedback can guide representation learning and exploration behavior [65], while sonic overlays can enrich the user experience of voice assistants beyond speech-only interaction [66]. More recent studies further demonstrate that dynamic natural soundscapes based on physiological data can support stress-related self-reflection and user experience [67], and that auditory sensations significantly affect human emotional responses in both subjective and physiological dimensions [68]. These findings suggest that non-speech auditory information can be used not only as a supplementary feedback signal but also as an active component of emotion-aware interaction design. Therefore, non-speech auditory cues could be integrated into the proposed framework either as an additional input modality for emotion inference or as an adaptive feedback channel for emotion-sensitive responses. For example, auditory overlays or affective sound cues may be dynamically adjusted according to the recognized emotional state of the user. Exploring the joint use of physiological, visual, and non-speech auditory information remains an important direction for future work. From a privacy perspective, the framework processes sensitive biometric data (e.g., EEG and facial signals) entirely on-device, without transmitting raw data to external servers. This design inherently reduces potential data leakage risks and supports privacy-preserving operation. Nevertheless, system-level considerations such as data retention policies, encryption mechanisms, user consent workflows, and compliance with privacy regulations (e.g., GDPR) are not fully addressed in this study. Future work will focus on large-scale and cross-cultural validation, comprehensive evaluation under diverse environmental and hardware conditions, and the integration of privacy-aware system-level mechanisms for real-world deployment.

5. Conclusions

The study enhances the adaptiveness of VUIs by recognizing users' emotional states from facial expressions and EEG signals. A multimodal dataset was collected using different VUI response types to elicit emotions in realistic interaction scenarios, enabling the development of a user-adaptive VUI that interprets emotional cues and adjusts its response style accordingly. Two emotion axes (stability and favorability) were derived from self-rated emotion scores to represent emotional coherence and evaluative preference. To support deployment on resource-constrained edge devices, the framework was optimized using CLR-RNF pruning and post-training quantization. Which significantly reduced model size and computational load while maintaining high accuracy. Future work will focus on full hardware deployment and real-world testing to realize real-time, emotionally intelligent VUIs for next-generation human–machine interaction.

CRediT authorship contribution statement

Van-Duc Khuat: Writing – original draft, Software. **Jeongin Kim:** Writing – original draft, Software. **Sang-Ho Kim:** Supervision, Conceptualization. **Yue Cao:** Investigation, Conceptualization. **Martin Maier:** Supervision, Conceptualization. **Wansu Lim:** Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Wansu Lim reports financial support was provided by National Research Foundation of Korea. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) (RS-2024-00349885), the Korea Health Industry Development Institute (KHIDI) (RS-2025-02223417), and the Korea Planning & Evaluation Institute of Industrial Technology (KEIT) (RS-2026-25517824).

Data availability

Data will be made available on request.

References

[1] B.-K. Ko, S.-H. Lee, S.-W. Lee, Imagined speech detection using multi-receptive CNN for asynchronous BCI communication and neurorehabilitation, IEEE Trans. Neural Syst. Rehabil. Eng. 33 (2025) 2904–2914, http://dx.doi.org/10.1109/TNSRE.2025.3592312.

[2] M. Norda, C. Engel, J. Rennies, J.-E. Appell, S.C. Lange, A. Hahn, Evaluating the efficiency of voice control as human machine interface in production, IEEE Trans. Autom. Sci. Eng. (2023) http://dx.doi.org/10.1109/TASE.2023.3302951, Early Access.

[3] D.-S. Kim, S.-H. Lee, K. Yin, S.-W. Lee, Reconstructing unseen sentences from speech-related biosignals for open-vocabulary neural communication, IEEE Trans. Neural Syst. Rehabil. Eng. 33 (2025) 4338–4348, http://dx.doi.org/10.1109/TNSRE.2025.3625219.

[4] S. Medjden, N. Ahmed, M. Lataifeh, Adaptive user interface design and analysis using emotion recognition through facial expressions and body posture from an RGB-D sensor, PLoS ONE 15 (7) (2020) http://dx.doi.org/10.1371/journal.pone.0235633.

[5] H. Cecotti, Adaptive time segment analysis for steady-state visual evoked potential based brain computer interfaces, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (3) (2020) 552–560, http://dx.doi.org/10.1109/TNSRE.2020.2968307.

[6] N. Rathnayake, D. Meedeniya, I. Perera, A. Welivita, A framework for adaptive user interface generation based on user behavioural patterns, in: Proc. MERCon, Moratuwa, Sri Lanka, 2019, pp. 698–703, http://dx.doi.org/10.1109/MERCon.2019.8818764.

[7] C. Nowacki, A. Gordeeva, A.H. Lizé, Improving the usability of voice user interfaces: A new set of ergonomic criteria, in: Proc. HCII, Copenhagen, Denmark, 2020, pp. 117–133, http://dx.doi.org/10.1007/978-3-030-50526-6_9.

[8] K. Seaborn, J. Urakami, Measuring voice UX quantitatively: A rapid review, in: Proc. CHI'21, Yokohama, Japan, 2021, pp. 1–8, http://dx.doi.org/10.1145/3411764.3445744.

[9] A.L. Iniguez-Carrillo, A. Venegas-Reynoso, L.S. Gaytán-Lugo, P.C. Santana-Mancilla, SOVO: Usability questionnaire for voice-only user interfaces, IEEE Lat. Am. Trans. 21 (11) (2023) 1164–1170, http://dx.doi.org/10.1109/TLA.2023.10268277.

[10] L. Shu, J. Xie, M. Yang, Z. Li, D. Liao, X. Xu, A review of emotion recognition using physiological signals, Sensors 18 (7) (2018) 20–74, http://dx.doi.org/10.3390/s18072074.

[11] D. Li, et al., Emotion recognition of subjects with hearing impairment based on fusion of facial expression and EEG topographic map, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 437–445, http://dx.doi.org/10.1109/TNSRE.2022.3225948.

[12] N. Saleem, et al., Attentional multimodal speech enhancement for voice user interface in consumer electronics, IEEE Trans. Consum. Electron. (2025) http://dx.doi.org/10.1109/TCE.2025.3603037, early access.

[13] Q. Zhao, et al., E-DANN: An enhanced domain adaptation network for audio-EEG feature decoupling in explainable depression recognition, IEEE Trans. Neural Syst. Rehabil. Eng. 33 (2025) 3647–3661, http://dx.doi.org/10.1109/TNSRE.2025.3608181.

[14] M. Chen, et al., Neural-free attention for monaural speech enhancement toward voice user interface for consumer electronics, IEEE Trans. Consum. Electron. 69 (4) (2023) 765–774, http://dx.doi.org/10.1109/TCE.2023.3254507.

[15] U. Côté-Allard, et al., Deep learning for electromyographic hand gesture signal classification using transfer learning, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (4) (2019) 760–771, http://dx.doi.org/10.1109/TNSRE.2019.2896269.

[16] D.H. Jin, Y.M. Kim, Y. Lee, Investigating the acceptance of voice user interfaces for users with disabilities, IEEE Access 13 (2025) 1055–1069, http://dx.doi.org/10.1109/ACCESS.2024.3520149.

[17] R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: Analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. 23 (10) (2001) 1175–1191, http://dx.doi.org/10.1109/34.954607.

[18] R.W. Picard, J. Healey, Affective wearables, in: Proc. ISWC, Cambridge, MA, USA, 1997, pp. 90–97, http://dx.doi.org/10.1109/ISWC.1997.629922.

[19] H. Chen, Y. Zhang, W. Liu, J. Wang, L. Huang, Quantifying emotional patterns for EEG-based emotion recognition: An interpretable study on EEG individual differences, IEEE Trans. Affect. Comput. http://dx.doi.org/10.1109/TAFFC.2025.3614727.

[20] S.S. Al-Hadithy, A.S. Abdalkafor, B. Al-Khateeb, Emotion recognition in EEG signals: Deep and machine learning approaches, challenges, and future directions, Comput. Biol. Med. 196 (2025) 110713.

[21] J. Li, J. Peng, End-to-end multimodal emotion recognition based on facial expressions and remote photoplethysmography signals, IEEE J. Biomed. Health Inform. 28 (10) (2024) 6054–6063, http://dx.doi.org/10.1109/JBHI.2024.3430310.

[22] Seong-Gyun Leem, et al., Selective acoustic feature enhancement for speech emotion recognition with noisy speech, IEEE/ACM Trans. Audio Speech Lang. Process. 32 (2023) 917–929.

[23] C.Y. Chang, J.S. Tsai, C.J. Wang, P.C. Chung, Emotion recognition with consideration of facial expression and physiological signals, in: Proc. IEEE CIBCB, Nashville, TN, USA, 2009, pp. 278–283, http://dx.doi.org/10.1109/CIBCB.2009.4938719.

[24] S. Dutta, S. Ganapathy, Leveraging content and acoustic representations for speech emotion recognition, IEEE Trans. Audio Speech Lang. Process. 33 (2025) 3678–3689, http://dx.doi.org/10.1109/TASLPRO.2025.3603853.

[25] X. Huang, J. Kortelainen, G. Zhao, X. Li, A. Moilanen, T. Seppänen, M. Pietikäinen, Multi-modal emotion analysis from facial expressions and electroencephalogram, Comput. Vis. Image Underst. 147 (2016) 114–124, http://dx.doi.org/10.1016/j.cviu.2015.09.015.

[26] J. Hu, Y. Huang, X. Hu, Y. Xu, The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses, IEEE Trans. Affect. Comput. 14 (1) (2023) 17–30, http://dx.doi.org/10.1109/TAFFC.2021.3118732.

[27] D. Swoboda, J. Boasen, P.-M. Léger, R. Pourchon, S. Sénécal, Comparing the effectiveness of speech and physiological features in explaining emotional responses during voice user interface interactions, Appl. Sci. 12 (3) (2022) 1269, http://dx.doi.org/10.3390/app12031269.

[28] M.B.H. Wiem, Z. Lachiri, Emotion assessing using valence-arousal evaluation based on peripheral physiological signals and support vector machine, in: Proc. CEIT, Hammamet, Tunisia, 2016, pp. 1–5, http://dx.doi.org/10.1109/CEIT.2016.7929097.

[29] S. Basu, N. Jana, A. Bag, M. Mahadevappa, J. Mukherjee, S. Kumar, R. Guha, Emotion recognition based on physiological signals using valence-arousal model, in: Proc. ICIIP, Shimla, India, 2015, pp. 50–55, http://dx.doi.org/10.1109/ICIIP.2015.7414763.

[30] M.B.H. Wiem, Z. Lachiri, Emotion sensing from physiological signals using three defined areas in arousal-valence model, in: Proc. ICCAD, Hammamet, Tunisia, 2017, pp. 219–223, http://dx.doi.org/10.1109/ICCAD.2017.8072711.

[31] X. Sun, Y. Zhang, J. Yang, et al., Current implications of EEG and fNIRS as functional neuroimaging tools: A comparative overview, Front. Neurosci. (2024) Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11629311/.

[32] W. Lin, H. Wang, J. Zhang, Review of studies on emotion recognition and judgment based on physiological signals, Appl. Sci. 13 (4) (2023) 2573, http://dx.doi.org/10.3390/app13042573.

[33] B. Rajasekhar, M. Kamaraju, V. Sumalatha, FPGA based emotions recognition from speech signals, in: Proc. ICBSII, Chennai, India, 2017, pp. 1–6, http://dx.doi.org/10.1109/ICBSII.2017.8072214.

[34] S. Saxena, S. Palaniswamy, S. Tripathi, Real-time emotion recognition from facial images using Raspberry Pi II, in: Proc. SPIN, Noida, India, 2016, pp. 666–670, http://dx.doi.org/10.1109/SPIN.2016.7566769.

[35] W.M. Ben Henia, Z. Lachiri, Embedded emotion recognition system based on electrocardiogram attributes, in: Proc. TSP, Athens, Greece, 2018, pp. 1–4, http://dx.doi.org/10.1109/TSP.2018.8441208.

[36] S. Dwijayanti, M. Iqbal, B.Y. Suprapto, Real-time implementation of face recognition and emotion recognition in a humanoid robot using a convolutional neural network, IEEE Access 10 (2022) 89876–89886, http://dx.doi.org/10.1109/ACCESS.2022.3212088.

[37] M. Lin, L. Cao, Y. Zhang, L. Shao, C.-W. Lin, R. Ji, Pruning networks with cross-layer ranking and k-reciprocal nearest filters, IEEE Trans. Neural Netw. Learn. Syst. 34 (11) (2023) 9139–9148, http://dx.doi.org/10.1109/TNNLS.2022.3156047.

[38] H. Cheng, M. Zhang, J.Q. Shi, A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations, IEEE Trans. Pattern Anal. Mach. Intell. 46 (12) (2024) 10558–10578, http://dx.doi.org/10.1109/TPAMI.2024.3447085.

[39] Y. Wang, Y. Qin, L. Liu, S. Wei, S. Yin, SWPU: A 126.04 TFLOPS/W edge-device sparse DNN training processor with dynamic sub-structured weight pruning, IEEE Trans. Circuits Syst. I: Regul. Pap. 69 (10) (2022) 4014–4027, http://dx.doi.org/10.1109/TCSI.2022.3184175.

[40] C.J. Schaefer, P. Taheri, M. Horeni, S. Joshi, The hardware impact of quantization and pruning for weights in spiking neural networks, IEEE Trans. Circuits Syst. II: Express Briefs 70 (5) (2023) 1789–1793, http://dx.doi.org/10.1109/TCSII.2023.3260701.

[41] B. Liu, T. Gu, H. Wang, Y. Qian, MixPQ: Joint pruning and quantization for speech and language foundation models compression, IEEE Trans. Audio Speech Lang. Process. http://dx.doi.org/10.1109/TASLPRO.2025.3613948.

[42] S. An, J. Shin, J. Kim, Quantization-aware training with dynamic and static pruning, IEEE Access 13 (2025) 57476–57484, http://dx.doi.org/10.1109/ACCESS.2025.3556629.

[43] Z. Wen, W. Lin, T. Wang, G. Xu, Distract your attention: Multi-head cross attention network for facial expression recognition, Biomimetics 8 (2) (2022) 199, http://dx.doi.org/10.3390/biomimetics8020199.

[44] R.T. Schirrmeister, et al., Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (2017) 5391–5420, http://dx.doi.org/10.1002/hbm.23730.

[45] Shaibal Saha, Lanyu Xu, Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies, Neurocomputing 643 (2025) 130417.

[46] Krishna Teja Chitty-Venkata, et al., A survey of techniques for optimizing transformer inference, J. Syst. Archit. 144 (2023) 102990.

[47] A. Keutayeva, B. Abibullaev, Data constraints and performance optimization for transformer-based models in EEG-based brain-computer interfaces: A survey, IEEE Access 12 (2024) 62628–62647.

[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. CVPR, Las Vegas, NV, USA, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[49] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, Ms-Celeb-1M: A dataset and benchmark for large-scale face recognition, in: Proc. ECCV, Amsterdam, The Netherlands, 2016, pp. 87–102, http://dx.doi.org/10.1007/978-3-319-46487-9_6.

[50] J.Y. Jeong, Y.G. Hong, D. Kim, J.W. Jeong, Y. Jung, S.H. Kim, Classification of facial expression in-the-wild based on ensemble of multi-head cross attention networks, in: Proc. IEEE CVPRW, New Orleans, LA, USA, 2022, pp. 2352–2357, http://dx.doi.org/10.1109/CVPRW56347.2022.00304.

[51] J.Y. Jeong, Y.G. Hong, S. Hong, J. Oh, Y. Jung, S.H. Kim, J.W. Jeong, Ensemble of multi-task learning networks for facial expression recognition in-the-wild with learning from synthetic data, in: Proc. ECCV, Tel Aviv, Israel, 2022, pp. 60–75, http://dx.doi.org/10.1007/978-3-031-25072-9_4.

[52] P. Ekman, Basic emotions, in: The Handbook of Cognition and Emotion, John Wiley & Sons Ltd., San Francisco, CA, USA, 1999, pp. 45–60.

[53] D. Kollias, ABAW: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges, in: Proc. CVPRW, New Orleans, LA, USA, 2022, pp. 2327–2335, http://dx.doi.org/10.1109/CVPRW56347.2022.00301.

[54] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. ICLR, San Diego, CA, USA, 2015, pp. 1–14, [Online]. Available: https://arxiv.org/abs/1409.1556.

[55] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: Proc. BMVC, Swansea, U.K., 2015, pp. 41.1–41.12, http://dx.doi.org/10.5244/C.29.41.

[56] S.I. Serengil, A. Ozpinar, Lightface: A hybrid deep face recognition framework, in: Proc. ASYU, Istanbul, Turkey, 2020, pp. 1–5, http://dx.doi.org/10.1109/ASYU50717.2020.9259802.

[57] S.I. Serengil, A. Ozpinar, Hyperextended Lightface: A facial attribute analysis framework, in: Proc. ICEET, Istanbul, Turkey, 2021, pp. 1–4, http://dx.doi.org/10.1109/ICEET53442.2021.9659815.

[58] P. Virtanen, et al., SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods 17 (2020) 261–272, http://dx.doi.org/10.1038/s41592-019-0686-2.

[59] A. Gramfort, et al., MEG and EEG data analysis with MNE-Python, Front. Neurosci. 7 (2013) http://dx.doi.org/10.3389/fnins.2013.00267.

[60] N. Kumar, K. Khaund, S.M. Hazarika, Bispectral analysis of EEG for emotion recognition, Procedia Comput. Sci. 84 (2016) 31–35, http://dx.doi.org/10.1016/j.procs.2016.04.062.

[61] S. Wu, X. Xu, L. Shu, B. Hu, Estimation of valence of emotion using two frontal EEG channels, in: Proc. BIBM, Kansas City, MO, USA, 2017, pp. 1127–1130, http://dx.doi.org/10.1109/BIBM.2017.8217803.

[62] J.G. Shin, I.G. Jo, W.S. Lim, S.H. Kim, A few critical design parameters affecting user's satisfaction in interaction with voice user interface of AI-infused systems, J. Ergon. Soc. Korea (JESK) 39 (1) (2020) 73–86, http://dx.doi.org/10.5143/JESK.2020.39.1.73.

[63] M. Nagamachi, Kansei engineering: The implication and applications to product development, in: Proc. IEEE SMC, Tokyo, Japan, 1999, pp. 273–278, http://dx.doi.org/10.1109/ICSMC.1999.814136.

[64] J.G. Shin, G.Y. Choi, H.J. Hwang, S.H. Kim, Evaluation of emotional satisfaction using questionnaires in voice-based human–AI interaction, Appl. Sci. 11 (4) (2021) 1920, http://dx.doi.org/10.3390/app11041920.

[65] X. Zhao, C. Weber, M.B. Hafez, S. Wermter, Impact makes a sound and sound makes an impact: Sound guides representations and explorations, in: Proc. IROS, Kyoto, Japan, 2022, pp. 2512–2518.

[66] M. Esau-Held, A. Marsh, V. Krauß, G. Stevens, Foggy sounds like nothing' Enriching the experience of voice assistants with sonic overlays, Pers. Ubiquitous Comput. 27 (2023) 1927–1947.

[67] R. Yan, X. Ren, S. Wang, X. Bai, X. Zhang, RainMind: Investigating dynamic natural soundscape of physiological data to promote self-reflection for stress management, Int. J. Hum.-Comput. Interact. 41 (9) (2025) 5545–5562.

[68] M. Yang, How auditory sensations affect human emotional responses to acoustic environments, J. Environ. Manag. 379 (2025) 124742.